

## **In-silico analysis of differentially expressed genes and their regulating microRNA involved in lymph node metastasis in invasive breast carcinoma**

Anupama Modi,<sup>1</sup> Purvi Purohit,\*<sup>1</sup> Ashita Gadwal,<sup>1</sup> Shweta Ukey,<sup>1</sup> Dipayan Roy,<sup>1</sup> Sujoy Fernandes,<sup>2</sup> Mithu Banerjee<sup>1</sup>

Author affiliations:

<sup>1</sup> Department of Biochemistry, All India Institute of Medical Sciences (AIIMS), Jodhpur, Rajasthan, India

<sup>2</sup> Department of Radiation Oncology, All India Institute of Medical Sciences (AIIMS), Jodhpur, Rajasthan, India

\*Corresponding author

Dr. Purvi Purohit

Additional Professor

Department of Biochemistry

AIIMS, Jodhpur

Rajasthan- 342005

E-mail: [dr.purvipurohit@gmail.com](mailto:dr.purvipurohit@gmail.com)

Phone: +91 9928388223

Funding: None.

## Abstract

**Introduction:** Axillary nodal metastasis is related to poor prognosis in breast cancer (BC). The metastatic progression in BC is related to molecular signatures. The currently popular methods to evaluate nodal status may give false negatives or give rise to secondary complications. In this study, key candidate genes in BC lymph node metastasis have been identified from publicly available microarray datasets and their roles in BC have been explored through survival analysis and target prediction.

**Methods:** Gene Expression Omnibus datasets have been analyzed for differentially expressed genes (DEGs) in lymph node-positive BC patients compared to nodal-negative and healthy tissues. The functional enrichment analysis was done in database for annotation, visualization and integrated discovery (DAVID). Protein-protein interaction (PPI) network was constructed in Search Tool for the Retrieval of Interacting Genes and proteins (STRING) and visualized on Cytoscape. The candidate hub genes were identified and their expression analyzed for overall survival (OS) in Gene Expression Profiling Interactive Analysis (GEPIA). The target miRNA and transcription factors were analyzed through miRNet.

**Results:** A total of 102 overlapping DEGs were found. Gene Ontology revealed eleven, seventeen, and three significant terms for cellular component, biological process, and molecular function respectively. Six candidate genes, DSC3, KRT5, KRT6B, KRT17, KRT81, and SERPINB5 were significantly associated with nodal metastasis and OS in BC patients. A total of 83 targeting miRNA were identified through miRNet and hsa-miR-155-5p was found to be the most significant miRNA which was targeting five out of six hub genes.

**Conclusion:** In-silico survival and expression analyses revealed six candidate genes and 83 miRNAs, which may be potential diagnostic markers and therapeutic targets in BC patients and miR-155-5p shows promise as it targeted five important hub genes related to lymph-node metastasis.

**Keywords:** Breast Cancer; lymph node metastasis; miRNA; miR-155-5p; in-silico

## 1. Introduction

According to the latest 2018 report of GLOBOCAN, breast cancer (BC) is the most common cancer in females which accounts for 11.6% among all cancer types. BC is commonly invasive, and there are 626,679 deaths due to BC worldwide, making it the second major cause of cancer-related deaths [1]. The involvement of axillary lymph nodes leads to worse prognosis in BC. According to the American Joint Committee on Cancer (AJCC)/International Union Against Cancer (UICC) tumor (T)-node (N)-metastasis (M) classification, nodal disease is classified in three groups based on the number of axillary metastatic lymph nodes involved: N1, 1–3 metastatic lymph node(s), N2, 4–9 metastatic lymph nodes, and N3, 10 or more metastatic lymph nodes [2]. The prognosis worsens with the increase in the number of metastatic lymph nodes [3, 4]. Five-year survival rates differ dramatically between women with negative lymph nodes (>90%) compared to those with lymph node metastasis (<70%) [5].

Positive or negative lymph node metastatic status aid in patient staging, prognostic information, and patient management [6]. Bakkour et al. observed that axillary lymph node involvement is significantly associated with overall survival (OS) and disease-free survival (DFS) [4]. Studies have also reported that a rise in the number of nodal involvement increases the rate of cancer recurrence [7, 8]. Although the involvement of axillary lymph node is the prognostic indicator of BC, the current method used to identify the nodal status, sentinel lymph node biopsy (SLNB), may give false-negative results, disrupt the lymphatic system and may lead to secondary complications. Therefore, there is a need for an appropriate method to identify the patients with and without lymph node metastasis, which could also reduce the chances of comorbidity related to surgical evaluation [9]. Identification of signature genes involved in lymph node metastasis of BC may be helpful in this process.

Differential gene expression signature can predict the tumor grade, subtypes and recurrence [9,10], which can predict the prognosis of diseases. Signature genes involved in local metastases in lymph node may not predict the outcome but it would have clinical and biological significance and improve our understanding of the process of metastasis [10]. Further, these genes could serve as therapeutic targets for the treatment of early-stage metastasis.

In this present study, we aimed to identify the crucial genes involved in lymph node metastasis in BC through publicly available microarray gene expression datasets available from the Gene Expression Omnibus (GEO). Through an in-silico approach, Kaplan-Meier survival analysis of hub genes was performed to identify the role of hub genes in the prognosis of BC. Further, for the hub genes which were significantly associated with poor prognosis of BC, their targeting miRNAs and transcription factors were identified.

## 2. Materials and Methods

### 2.1 Data collection

Gene expression profiles of axillary lymph node (N<sup>-</sup>) (GSE42568, GSE22093, GSE76275, GSE23988, and GSE36771) and non-axillary lymph node involving (N<sup>+</sup>) BC tissue, and healthy breast tissue (GSE42568) were downloaded from public database GEO [11-13]. These datasets were classified into three groups: 1) N<sup>-</sup> vs. control tissues (GSE42568) which included 45 N<sup>-</sup> BC tissues and 17 control breast tissues, 2) N<sup>+</sup> vs. control tissues (GSE42568) contained 59 N<sup>+</sup> BC tissues and 17 control breast tissues, and 3) N<sup>-</sup> vs. N<sup>+</sup> (GSE42568, GSE22093, GSE76275, GSE23988, and GSE36771) including a total of 219 N<sup>-</sup> and 250 N<sup>+</sup> BC tissues, respectively (Table 1).

### 2.2 Identification of differentially expressed genes (DEGs)

The interactive web tool GEO2R (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>) was used to compare between groups of samples for differential expression analysis. The DEGs between N<sup>-</sup> vs. Control, N<sup>+</sup> vs. Control and N<sup>-</sup> vs. N<sup>+</sup> groups from GSE datasets were determined according to the cut-off value of  $|\log_2\text{fold change (FC)}| \geq 0.58$  and  $P\text{-value} < 0.05$ . For visualization of DEGs, volcano plot was constructed for each dataset using R [14] (Figure 2a-g). Venn diagram was constructed using Venn diagram tool available at <http://bioinformatics.psb.ugent.be/webtools/Venn/> for identifying the overlapping DEGs. Workflow for processing of data has been shown in figure 1.

### 2.3 Functional Enrichment of DEGs

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway functional enrichment were analyzed through the database for annotation, visualization and integrated discovery (DAVID; <https://david.ncifcrf.gov/>) [15].  $P\text{-value} < 0.05$  was considered to be statistically significant. For the overlapping DEGs, GO enrichment

analyses were classified into three functional groups: biological processes (BP), cellular components (CC), and molecular functions (MF).

#### *2.4 PPI network construction*

Protein-protein Interaction (PPI) network of overlapping 102 DEGs was constructed using the Search Tool for the Retrieval of Interacting Genes and proteins (STRING) database (<https://string-db.org/>) [16] which analyse the functional interaction between proteins. To explore the regulatory mechanisms, interactions with the confidence of a combined score  $>0.400$  were retained and imported to Cytoscape (version 3.8.0) [17] for visualization of the PPI network.

#### *2.5 Hub gene analysis*

In Cytoscape, the CytoHubba plug-in, which uses 12 different methods and provides a user-friendly interface to analyze the topology of PPI networks to select the top 10 genes, was employed [18]. Those genes which were detected with at least three different methods were considered as the hub genes.

#### *2.6 Expression and survival analysis of hub genes*

Gene Expression Profiling Interactive Analysis (GEPIA) (<http://gepia.cancer-pku.cn/>) [19] is an interactive web server, which analyzes RNA-sequence expression data of 9,736 tumors and 8,587 normal samples from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) projects. It can also conduct OS by log-rank test based on relative gene expression. The Cox proportional hazard ratio (HR) and the 95% confidence interval (95% CI) of the survival plot can also be obtained. In our study, we explored the OS of individual hub genes through the Kaplan–Meier plotter and tissue expression of hub genes in BC.

#### *2.7 Hub gene targeting miRNA and transcription factor analysis*

Hub gene targeting miRNAs were predicted by using miRNet 2.0 database (<https://www.mirnet.ca/miRNet/home.xhtml>), a miRNA-centric network visual analytics platform, and hub gene targeting transcription factors were analysed using miRNet at five different platforms (CHEA, ENCODE, JASPAR, REGNETWOTK and TRRUST). Further, hub gene-miRNA and hub gene-transcription factor interaction networks were visualized in Cytoscape.

### 3. Results

#### 3.1 Identification of DEGs

For the chosen datasets, the number of DEGs were 7935 (GSE42568, N<sup>-</sup> vs. control), 8298 (GSE42568, N<sup>+</sup> vs. control), 221 (GSE42568, N<sup>-</sup> vs. N<sup>+</sup>), 292 (GSE76275), 333 (GSE23988), and 551 (GSE22093), respectively. The Venn diagram for DEGs was constructed which showed 949 genes in the N<sup>+</sup> group. To identify the genes differentially expressed in N<sup>+</sup> vs. N<sup>-</sup> group, from data sets (GSE42568, GSE22093, GSE76275, GSE23988 and GSE 36771) DEGs were downloaded. These DEGs were combined, and after removing the duplicate genes, there were 1287 DEGs. For further analysis, Venn diagram of 949 genes which were differentially expressed in N<sup>+</sup> BC compared to controls, and 1287 DEGs of N<sup>+</sup> vs. N<sup>-</sup> group was constructed, which revealed that there were 102 overlapping DEGs based on the cut-off criteria of  $|\log_2FC| \geq 0.58$  and  $P$  value < 0.05 (Figure 3).

#### 3.2 Functional enrichment analysis of DEGs

Three GO terms: CC, BP, and MF were analysed. KEGG pathways were analysed to identify the pathways related to the DEGs. In CC, BP, and MF, eleven, seventeen, and three GO terms, respectively, were found to be statistically significant. For CC, most of the genes were involved in Immunoglobulin complex, circulating; Blood microparticles; External side of plasma membrane; Extracellular exosome; Extracellular space; whereas for BP, there was Positive regulation of B-cell activation; Phagocytosis, recognition; Phagocytosis, engulfment; B cell receptor signalling pathway; Complement activation, classical pathway. For MF, the significant terms were Immunoglobulin receptor binding, Antigen binding, and Serine-type endopeptidase activity. Similarly, KEGG pathway enrichment analysis revealed that the DEGs were mainly involved in the p53 signalling pathway and Aldosterone synthesis and secretion (Figure 4a-d, Table 2-5).

#### 3.3 PPI network construction and module analysis

To identify the interactions between the overlapping DEGs, online database STRING was used. A total of 86 DEGs as network nodes and 244 edges were used to construct and visualize the PPI network in Cytoscape (Figure 5).

#### 3.4 Hub gene analysis

Hub genes were identified using CytoHubba application, in which overlapping top 10

genes observed using twelve topological analysis methods, revealed that there were 21 hub genes, among which 13 genes were identified by at least three different methods as the candidate hub genes (Table 6).

### *3.5 Survival analysis*

Association between hub gene expressions with OS of BC patients were analysed to examine the role of hub genes in BC prognosis by grouping the population into high expression and low expression categories, six genes showed a significant association between differential expressions with shorter OS among patients with BC. These six hub genes- DCS3 (HR=0.72, p=0.045), KRT17 (HR=0.51, p=0.002), KRT5 (HR=0.71, p=0.035), KRT6B (HR=0.57, p<0.001), KRT81 (HR=0.70, p=0.029) and SERPINB5 (HR=0.68, p=0.017) may serve as prognostic biomarkers to determine the severity of BC patients (Figure 6a-f). The expressions of the six hub genes in BC tissue compared with normal tissue are shown in Figure 7a-f.

### *3.6 Hub genes targeting transcription factors and miRNA analysis*

Hub genes targeting miRNA were predicted using miRNet, based on the correlation analysis between the hub genes and miRNAs. In the network, there were 6 genes and 83 miRNA, with 6 nodes and 110 edges. Out of these 83 miRNAs, hsa-miR-155-5p targeted all hub genes except KRT81. Further, using five different databases, gene-transcription factor network was analyzed, as depicted in Figure 8.

## **4. Discussion**

The current study aimed to gain insight into the change in gene expression in LNM in BC through analyzing publicly available datasets. For this, we have identified the overlapping DEGs responsible for LNM in BC. Among these DEGs, six candidate hub genes, namely DSC3, KRT17, KRT5, KRT6B, KRT81 and SERPINB5, showed significant associations with poor OS in BC patients. Further, hub gene targeting miRNA were also identified. A total of 83 miRNA and 112 transcription factors were identified for these selected genes.

Axillary lymph node status is an important prognostic factor in BC. An increase in the number of metastatic lymph nodes involved is associated with a decrease OS as well as DFS [5,20]. Axillary lymph node status is also pivotal in determining the course of management e.g., radiotherapy in these patients [21,22]. Accordingly, the better

prognosis of node-negative patients of BC is attributed to timely resection, before distant metastasis via the axillary lymphatics has occurred [23]. The numbers of these metastatic lymph nodes are those that are dissected by the surgeon and examined by the pathologist. However, various studies have shown that the metastatic lymph nodes are greater in number with an increasing number of removed nodes. Thus, it is difficult to assess the axillary lymph node status reliably without removing and identifying sufficient numbers of lymph nodes depending on the surgeon and pathologist [24-26]. Further, nodal status identification involves lymph node biopsy, which can potentially give false-negative results. Therefore, there is a need for an appropriate method to identify patients with and without LNM, which could also reduce the chances of comorbidity related to surgical evaluation. Identification of candidate genes related to nodal metastasis can aid in the better evaluation and subsequent management of these patients. Furthermore, BC metastasis has been known to be molecularly distinct from their primary tumor counterparts.

Desmocollin 3 (DSC3), a member of the cadherin superfamily of calcium-dependent cell adhesion molecules, is a desmosomal protein. DSC3 helps in maintaining tissue architecture; hence, their loss leads to a lack of adhesion and a gain of cellular mobility [27,28]. DSC3 is a p53 responsive gene, and its expression is down-regulated in BC cell lines and primary breast tumors, indicating that the loss of DSC3 expression is a common event in primary breast tumor specimens [27]. In esophageal adenocarcinoma (EACs) tissue samples and human EAC cell lines, a significant down-regulation ( $P<0.001$ ) of the DSC3 mRNA levels has been observed. In addition, the EAC cell lines and tumor samples had aberrant promoter hypermethylation as compared to normal esophageal samples ( $P<0.001$ ) [28]. DSC3 has also been implicated in LNM and cellular proliferation in oral squamous cell carcinoma (OSCC) through the regulation of  $\beta$ -catenin [29].

KRT genes are involved in the synthesis of keratin, which is a fibrous protein involved in the structure of epithelial cells. KRT5 is overexpressed in basal-like BC subtype tissue, as well as cell line and is associated with poor outcome [30]. It is also seen to be overexpressed in younger women with BC compared to older women [31] KRT6B, a key mediator of notch signalling [32], has been observed to be overexpressed in HCC and



responsible for its progression. In BC, it is a favourable prognostic marker. Keratin 17 (KRT17) is a 48KDa type I intermediate filament, which is mainly expressed in epithelial basal cells. KRT17 is overexpressed in many malignant tumors and plays an important role in tumor development. It has been observed that KRT17 expression levels were significantly higher in lung cancer compared to normal lung tissues, and such a high expression of KRT17 predicted poor prognosis for patients with lung adenocarcinomas and was correlated with poor differentiation and lymphatic metastasis [33].

Overexpression of KRT17 enhanced, while its knockdown inhibited, the proliferation and invasiveness of lung cancer cells [33]. KRT17 promotes ESCC cell proliferation and migration, thus potentially contributing to metastasis. It also induces epithelial-mesenchymal transition through activation of AKT signalling [34]. Both KRT5 and KRT17 are positive markers for TNBCs [35]. KRT81 is found to be higher in BC tissue compared with normal mammary epithelial cells. In KRT81-knockdown MDA-MB231 cells, a decreased MMP9 activity was noticed, while along with decreased cell migration and invasion capabilities [36].

High immunoreactivity of SERPINB5 or Maspin (mammary serine protease inhibitor), is significantly linked to pre and post-CCRT advanced disease, lymphovascular invasion, and poor response to colorectal cancer (all  $P \leq 0.015$ ). It is a metastasis suppressor gene. SERPINB5 overexpression is negatively associated with disease-specific survival (DSS), local recurrence-free survival (LRFS) and metastasis-free survival (MeFS) rates in colorectal cancer. It can also independently predict for DSS and MeFS in colorectal cancer (all  $P \leq 0.043$ ) [37]. In fact, SERPINB5 is part of the metastatic epithelial-gene signature in BC and associated with an unfavourable prognosis [38]. In BC cell line, SERPINB5 inhibits tumour cell invasion and promotes cellular adhesion [39].

In this study, we also identified the miRNAs targeting these hub genes and hsa-miR155-5p was identified as targeting five hub genes. Similar findings have been reported in a clinical study involving 80 individuals, where BC patients showed a significant correlation with lymph node positive status and miR-155-5p expression [40]. Further, in-silico survival analysis found the genes to be significantly associated with OS, suggesting that these hub genes may serve as potential prognostic biomarkers and therapeutic targets for BC. However, there were certain limitations to our study. Firstly, when analyzing the

DEGs, given the complexity of datasets in our study, it is difficult to consider demographic factors such as different age groups, ethnicity, geographical regions, as well as tumor staging and classification of all the patients. Secondly, according to the results, the six hub genes were all up or downregulated in BC, but the mechanism of their dysregulation is yet to be explored. Therefore, more pieces of evidences are required to find out their mechanistic foundation. The hsa-miR155-5p identified as common miR regulating the expression of hub genes needs further exploration in regards to its role in LNM. The transcription factors identified in this study could aid in the exploration of those pathways. Finally, this study has analyzed the expression levels and OS of the six hub genes with the help of bioinformatics, which may help to develop a biomarker panel for this patient population. However, to establish these hub genes as diagnostic or prognostic markers with high accuracy and specificity for BC, we need larger, prospective studies.

### **List of abbreviations:**

BC: Breast Cancer; CCRT: Concurrent chemoradiotherapy; CHEA: Chip Enrichment Analysis; DAVID: database for annotation, visualization and integrated discovery; DEG: Differentially expressed genes; DFS: Disease-free survival; DSS: Disease specific survival; ENCODE: Encyclopedia of DNA elements; GEO: Gene Expression Omnibus; GEPIA: Gene Expression Profiling Interactive Analysis; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; LNM: Lymph Node Metastasis; LRFS: Local recurrence-free survival; Maspin: mammary serine protease inhibitor; MeFS: metastasis free survival; miRNA: MicroRNA; OS: Overall Survival; PPI: Protein-protein interaction; SLNB: sentinel lymph node biopsy; STRING: Search Tool for the Retrieval of Interacting Genes and proteins; TRRUST: Transcriptional Regulatory Relationships Unravalled by Sentence-based Text Mining

**Conflict of interest:** The authors declare no conflicts of interest.

**Acknowledgement:** The authors acknowledge the support of the Department of Biotechnology, Ministry of Science and Technology (no. DBT/2018/AIMS-J/994).

## References

- [1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* [Internet]. 2018 Nov [cited 2020 Apr 12];68(6):394–424. Available from: <http://doi.wiley.com/10.3322/caac.21492>
- [2] Edge SB, Byrd DR, Compton CC et al., *AJCC Cancer Staging Handbook*, Springer, New York, NY, USA, 7th edition, 2010.
- [3] Fisher B, Bauer M, and Wickerham DL, “Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update,” *Cancer*, vol. 52, no. 9, pp. 1551–7, 1983.
- [4] BAKKOUR, Amine Mohammed et al. The predictors and the prognostic significance of axillary lymph nodes involvement in breast cancer. *International Surgery Journal*, [S.l.], v. 6, n. 5, p. 1641-5, apr. 2019. ISSN 2349-902. doi: 10.18203/2349-2902.isj20191512.
- [5] Eifel P, Axelson JA, Costa J et al., “National institutes of health consensus development conference statement: adjuvant therapy for breast cancer, November 1–3, 2000,” *Journal of the National Cancer Institute*, vol. 93, no. 13, pp. 979–89, 2001.
- [6] Pinheiro DJ, Elias S, Nazário AC. Axillary lymph nodes in breast cancer patients: sonographic evaluation. *Radiol Bras.* 2014;47(4):240-4. doi:10.1590/0100-3984.2013.1689
- [7] Vinh-Hung V, Cserni G, Burzykowski T, Van de Steene J, Voordeckers M, Storme G. Effect of the number of uninvolved nodes on survival in early breast cancer. *Oncol Reports.* 2003;10(2):363-8.
- [8] Arvold ND, Taghian AG, Niemierko A, et al. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. *J Clin Oncol.* 2011;29(29):3885.

- [9] Ellsworth RE, Field LA, Love B, Kane JL, Hooke JA, Shriver CD. Differential gene expression in primary breast tumors associated with lymph node metastasis. *Int J Breast Cancer*. 2011;2011:142763. doi:10.4061/2011/142763
- [10] van't Veer LJ, Dai H, van de Vijver MJ et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–6, 2002.
- [11] Clarke C, Madden SF, Doolan P, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* 2013 Oct;34(10):2300-8. PMID: 23740839
- [12] Burstein MD, Tsimelzon A, Poage GM, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 2015 Apr 1;21(7):1688-98. PMID: 25208879
- [13] Iwamoto T, Bianchini G, Booser D, Qi Y et al. Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *J Natl Cancer Inst* 2011 Feb 2;103(3):264-72. PMID: 21191116
- [14] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [15] Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4: P3, 2003.
- [16] Szklarczyk D, Morris JH, Cook H, et al: The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45: D362-D368, 2017
- [17] Shannon P, Markiel A, Ozier O et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–504, 2003.

- [18] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT and Lin CY: CytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC SystBiol* 4 (Suppl 8): S11, 2014.
- [19] Tang Z, Li C, Kang B, Gao G, Li C, and Zhang Z, “GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses,” *Nucleic Acids Research*, vol. 45, no. 1, pp. W98–W102, 2017.
- [20] Fisher B, Bauer M, and Wickerham DL, “Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update,” *Cancer*, vol. 52, no. 9, pp. 1551–7, 1983.
- [21] Krag DN, Anderson SJ, Julian TB, et al. Technical outcomes of sentinel-lymph-node resection and conventional axillary-lymph-node dissection in patients with clinically node-negative breast cancer: results from the NSABP B-32 randomised phase III trial. *The Lancet Oncol.* 2007;8(10):881-8.
- [22] Giuliano AE, et al., Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial, *Jama* 305 (6) (2011) 569–75.
- [23] Eggers C: Cancer surgery: The value of radical operations for cancer after the lymphatic drainage area has become involved. *Ann Surg* 106:668-9, 1937
- [24] Schaapveld M, Otter R, De Vries EGE et al., “Variability in axillary lymph node dissection for breast cancer,” *Journal of Surgical Oncology*, vol. 87, no. 1, pp. 4–12, 2004.
- [25] Fortin A, Dagnault A, Blondeau L, Vu TTT, and Larochelle M, “The impact of the number of excised axillary nodes and of the percentage of involved nodes on regional nodal failure in patients treated by breast-conserving surgery with or without regional irradiation,” *International Journal of Radiation Oncology Biology Physics*, vol. 65, no. 1, pp. 33–9, 2006.

- [26] Cil T, Hauspy J, Kahn H et al., “Factors affecting axillary lymph node retrieval and assessment in breast cancer patients,” *Annals of Surgical Oncology*, vol. 15, no. 12, pp. 3361–8, 2008.
- [27] Oshiro MM, Kim CJ, Wozniak RJ, et al. Epigenetic silencing of DSC3 is a common event in human breast cancer. *Breast Cancer Res.* 2005;7(5):R669-80. doi: 10.1186/bcr1273. Epub 2005 Jun 16. PMID: 16168112; PMCID: PMC1242132.
- [28] Wang Q, Peng D, Zhu S, et al. Regulation of Desmocollin3 Expression by Promoter Hypermethylation is Associated with Advanced Esophageal Adenocarcinomas. *J Cancer.* 2014;5(6):457-64. Published 2014 May 14. doi:10.7150/jca.9145
- [29] Wang L, Liu T, Wang Y, et al. Altered expression of desmocollin 3, desmoglein 3, and beta-catenin in oral squamous cell carcinoma: correlation with lymph node metastasis and cell proliferation. *Virchows Arch.* 2007 Nov;451(5):959-66. doi: 10.1007/s00428-007-0485-5.
- [30] Kuo WH, Chang YY, Lai LC, et al. Molecular characteristics and metastasis predictor genes of triple-negative breast cancer: a clinical study of triple-negative breast carcinomas. *PLoS One.* 2012;7(9):e45831. doi: 10.1371/journal.pone.0045831. Epub 2012 Sep 25. PMID: 23049873; PMCID: PMC3458056.
- [31] Johnson RH, Hu P, Fan C, Anders CK. Gene expression in "young adult type" breast cancer: a retrospective analysis. *Oncotarget.* 2015 May 30;6(15):13688-702. doi: 10.18632/oncotarget.4051. PMID: 25999348; PMCID: PMC4537042.
- [32] Zhang H, Huo M, Jia Y, Xu A. KRT6B, a key mediator of notch signaling in honokiol-induced human hepatoma cell apoptosis. *Int J Clin Exp Med.* 2015 Sep 15;8(9):16880-9. PMID: 26629239; PMCID: PMC4659127.
- [33] Wang Z, Yang MQ, Lei L, et al. Overexpression of KRT17 promotes proliferation and invasion of non-small cell lung cancer and indicates poor prognosis. *Cancer Manag Res.* 2019;11:7485-97. Published 2019 Aug 7. doi:10.2147/CMAR.S218926

- [34] Liu Z, Yu S, Ye S, et al. Keratin 17 activates AKT signalling and induces epithelial-mesenchymal transition in oesophageal squamous cell carcinoma. *J Proteomics*. 2020 Jan 16;211:103557. doi: 10.1016/j.jprot.2019.103557.
- [35] Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol*. 2016 Nov;13(11):674-90. doi: 10.1038/nrclinonc.2016.66. Epub 2016 May 17. PMID: 27184417; PMCID: PMC5461122.
- [36] Nanashima N, Horie K, Yamada T, Shimizu T, Tsuchida S. Hair keratin KRT81 is expressed in normal and breast cancer cells and contributes to their invasiveness. *Oncol Rep*. 2017 May;37(5):2964-70. doi: 10.3892/or.2017.5564. Epub 2017 Apr 11. PMID: 28405679.
- [37] Chang IW, Liu KW, Ragunanan M, He HL, Shiue YL, Yu SC. SERPINB5 Expression: Association with CCRT Response and Prognostic Value in Rectal Cancer. *Int J Med Sci*. 2018 Feb 12;15(4):376-84. doi: 10.7150/ijms.22823. PMID: 29511373; PMCID: PMC5835708.
- [38] Vecchi M, Confalonieri S, Nuciforo P. Breast cancer metastases are molecularly distinct from their primary tumors. *Oncogene*. 2008 Apr 3;27(15):2148-58. doi: 10.1038/sj.onc.1210858. Epub 2007 Oct 22. PMID: 17952122.
- [39] Seftor RE, Seftor EA, Sheng S, Pemberton PA, Sager R, Hendrix MJ. Maspin suppresses the invasive phenotype of human breast carcinoma. *Cancer Res*. 1998; 58: 5681-5.
- [40] Chernyy V, Pustylnyak V, Kozlov V, Gulyaeva L. Increased expression of miR-155 and miR-222 is associated with lymph node positive status. *J Cancer*. 2018;9(1):135-140. Published 2018 Jan 1. doi:10.7150/jca.22181



## Table and Figure Legends

**Table 1:** Group-wise distribution of breast cancer tissue samples obtained from each dataset

**Table 2:** Gene Ontology analysis for the overlapping genes ~ Biological Process

**Table 3:** Gene Ontology analysis for the overlapping genes ~ Molecular Function

**Table 4:** Gene Ontology analysis for the overlapping genes ~ Cellular Components

**Table 5:** KEGG pathway analysis for the overlapping genes

**Table 6:** Hub gene analysis using CytoHubba plug-in to identify the key candidate genes

**Table 7:** Common miRNA targeting hub genes

**Figure 1:** A schematic diagram of the workflow

**Figure 2a-g:** Volcano plots for the differentially expressed genes (DEGs) for each dataset

**Figure 3:** Venn diagram depicting the 102 overlapping DEGs for N<sup>+</sup> BC patients

**Figure 4a-d:** Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis by DAVID

**Figure 5:** The protein-protein interaction (PPI) network for the selected hub genes visualized on Cytoscape

**Figure 6a-f:** Associations of the six candidate hub genes with overall survival in breast cancer.

**Figure 7a-f:** Differential expressions of the six candidate hub genes in breast cancer.

**Figure 8:** The hub gene-miRNA-transcription factor regulatory network

**Table 1:** Group-wise distribution of breast cancer tissue sample obtained from each dataset

S.No.	GEO Dataset	No. of control samples	No. of N <sup>+</sup> samples	No. of N <sup>-</sup> samples
1.	GSE42568	17	59	45
2.	GSE23988		40	21
3.	GSE22093		30	21
4.	GSE76275		76	74
5.	GSE36771		45	58
	Total	17	250	219

**Table 2:** Gene Ontology analysis for the overlapping genes ~ Biological Process

Term	%	p value	Genes
GO:0050871~positive regulation of B cell activation	7.920792	0.000	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2
GO:0006910~phagocytosis, recognition	7.920792	0.000	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2
GO:0006911~phagocytosis, engulfment	7.920792	0.000	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2
GO:0050853~B cell receptor signaling pathway	7.920792	0.000	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2
GO:0006958~complement activation, classical pathway	8.910891	0.000	IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2
GO:0001895~retina homeostasis	4.950495	0.000	IGHG3, IGKC, PIP, IGHA1, IGHA2
GO:0006956~complement activation	5.940594	0.000	IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, IGLC1
GO:0006955~immune response	9.90099	0.000	IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, TRBC1, IGHD, IGLC1, IGHA1, IGHA2, ILF2
GO:0045087~innate immune response	9.90099	0.000	IGHG3, IGHG4, IGKC, CRISP3, IGHV3-23, TMED7-TICAM2, IGHD, IGLC1, IGHA1, IGHA2
GO:0060267~positive regulation of respiratory burst	2.970297	0.000	CAMK1D, IGHA1, IGHA2
GO:0006898~receptor-mediated endocytosis	6.930693	0.000	MSR1, IGKC, IGLV1-44, IGHV3-23, IGLC1, IGHA1, IGHA2
GO:0038096~Fc-gamma receptor signaling pathway involved in phagocytosis	5.940594	0.000	IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, IGLC1
GO:0042742~defense response to	5.940594	0.001	IGHG3, IGHG4, IGKC, IGHV3-23,

bacterium			IGHD, IGLC1
GO:0050776~regulation of immune response	5.940594	0.002	IGKC, IGLV1-44, IGHV3-23, TRBC1, IGLC1, PVR, IGHG3, IGHG4, IGKC, CAPN6, PIP, IGLV1-44, IGHV3-23, PREP,
GO:0006508~proteolysis	8.910891	0.004	IGLC1
GO:1901621~negative regulation of smoothened signaling pathway involved in dorsal/ventral neural tube patterning	1.980198	0.030	GPR161, PRKACB
GO:0003094~glomerular filtration	1.980198	0.045	IGHA1, IGHA2
GO:0042989~sequestering of actin monomers	1.980198	0.050	TMSB15A, TMSB15B
GO:0038095~Fc-epsilon receptor signaling pathway	3.960396	0.063	IGKC, IGLV1-44, IGHV3-23, IGLC1
GO:0018298~protein-chromophore linkage	1.980198	0.079	OPN3, IGHA1
GO:0050871~positive regulation of B cell activation	7.920792	0.000	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2

**Table 3:** Gene Ontology analysis for the overlapping genes ~ Molecular Function

Term	%	p value	Genes
GO:0034987~immunoglobulin receptor binding	7.920792	0.00	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2, IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, IGHD, IGLC1, IGHA1,
GO:0003823~antigen binding	8.910891	0.00	IGHA2
GO:0004252~serine-type endopeptidase activity	6.930693	0.00	IGHG3, IGHG4, IGKC, IGLV1-44, IGHV3-23, PREP, IGLC1
GO:0016491~oxidoreductase activity	3.960396	0.08	CYB5R2, SESN3, HSD17B6, FAR2
GO:0001965~G-protein alpha-subunit binding	1.980198	0.09	RGS1, F2R

**Table 4:** Gene Ontology analysis for the overlapping genes ~ Cellular Components

Term	%	p value	Genes
GO:0042571~immunoglobulin complex, circulating	5.940594	0.00	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1
GO:0072562~blood microparticle	8.910891	0.00	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2, CP
GO:0009897~external side of plasma membrane	7.920792	0.00	IGHG3, IGHG4, IGKC, IGHV3-23, IGHD, IGLC1, IGHA1, IGHA2, STEAP4, CLIC3, CRISP3, IGHV3-23, PVR, IGHG3, IGHG4, MIEN1, IGKC, HIST1H2AD, FLNA, IGLC1, PROM1, IGHA1, IGHA2, PRKACB, S100A11, KRT6B, TMED7-TICAM2, KRT5, CP, SERPINB5, TOM1L1, KRT17, PIP,
GO:0070062~extracellular exosome	26.73267	0.00	IGHD, ZNF711, KRT81, CRISP3, IGHV3-23, CP, PVR, SERPINB5, IGHG3, IGHG4, IGKC, PIP, IGLC1, IGHA1, PROM1, IGHA2,
GO:0005615~extracellular space	14.85149	0.01	S100A11
GO:0031941~filamentous actin	2.970297	0.01	TMSB15A, TMSB15B, FLNA, CNTNAP3, CRISP3, F2R, IGHV3-23, CP, IGHG3, IGHG4, IGKC, PIP,
GO:0005576~extracellular region	15.84158	0.01	IGLV1-44, FLNA, IGLC1, COL9A3, IGHA1, IGHA2, DSC3
GO:0071751~secretory IgA immunoglobulin complex	1.980198	0.01	IGHA1, IGHA2
GO:0071752~secretory dimeric IgA immunoglobulin complex	1.980198	0.01	IGHA1, IGHA2

---

GO:0071748~monomeric IgA			
immunoglobulin complex	1.980198	0.01	IGHA1, IGHA2
GO:0005793~endoplasmic reticulum-Golgi intermediate compartment			
compartment	2.970297	0.04	TMED7-TICAM2, PROM1, CNIH4
GO:0045095~keratin filament			
keratin filament	2.970297	0.09	KRT81, KRT5, KRT6B

---

**Table 5:** KEGG pathway analysis for the overlapping genes

Term	%	p value	Genes
hsa04115:p53 signaling pathway	2.970297	0.032	SESN3, CDKN2A, SERPINB5
hsa04925:Aldosterone synthesis and secretion	2.970297	0.045	PLCB4, CAMK1D, PRKACB
hsa04540:Gap junction	2.970297	0.053	PLCB4, PRKACB, SOS2
hsa04912:GnRH signaling pathway	2.970297	0.056	PLCB4, PRKACB, SOS2
hsa04915:Estrogen signaling pathway	2.970297	0.065	PLCB4, PRKACB, SOS2 PLCB4, CDKN2A, F2R, PRKACB,
hsa05200:Pathways in cancer	4.950495	0.081	SOS2

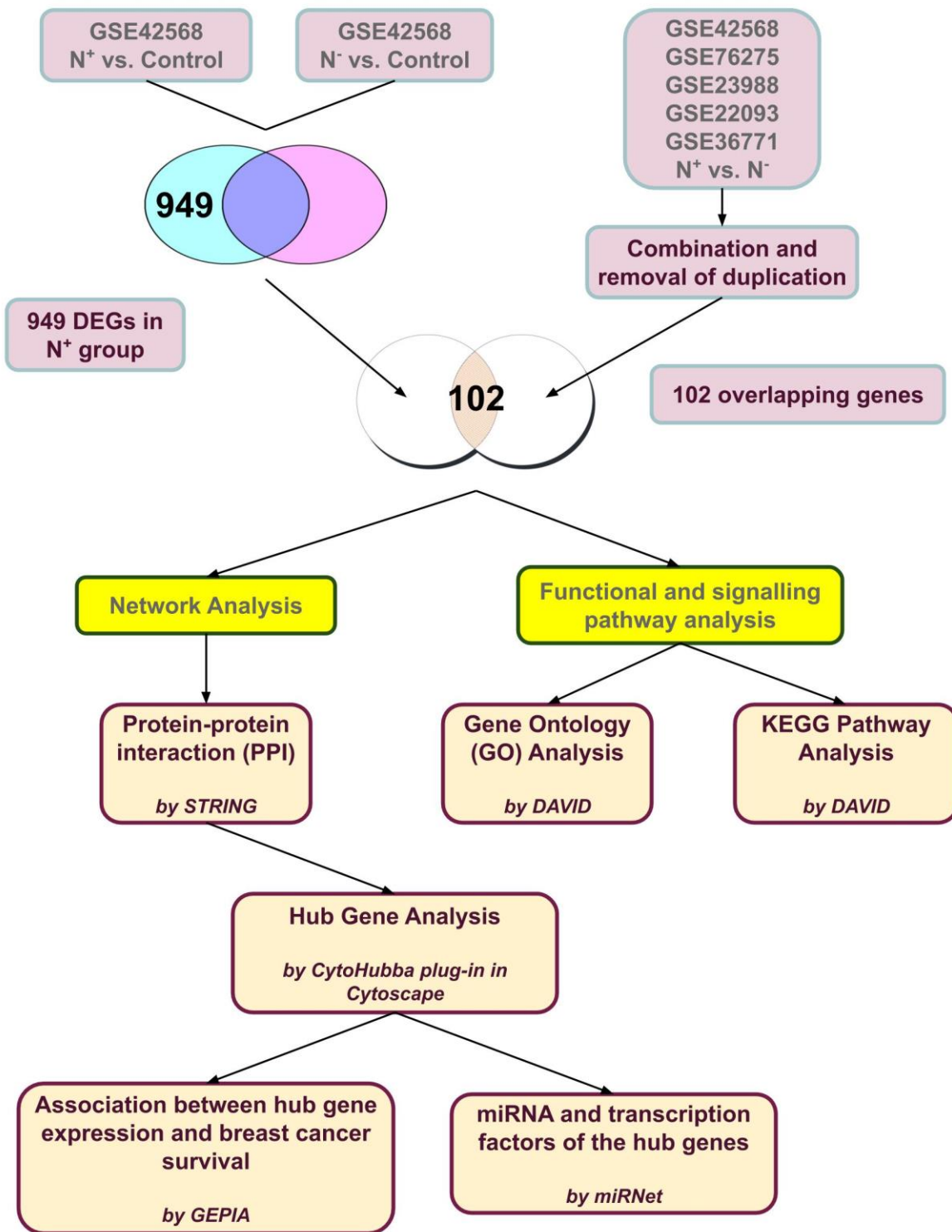


**Table 6:** Hub gene analysis using CytoHubba plug-in to identify the key candidate genes

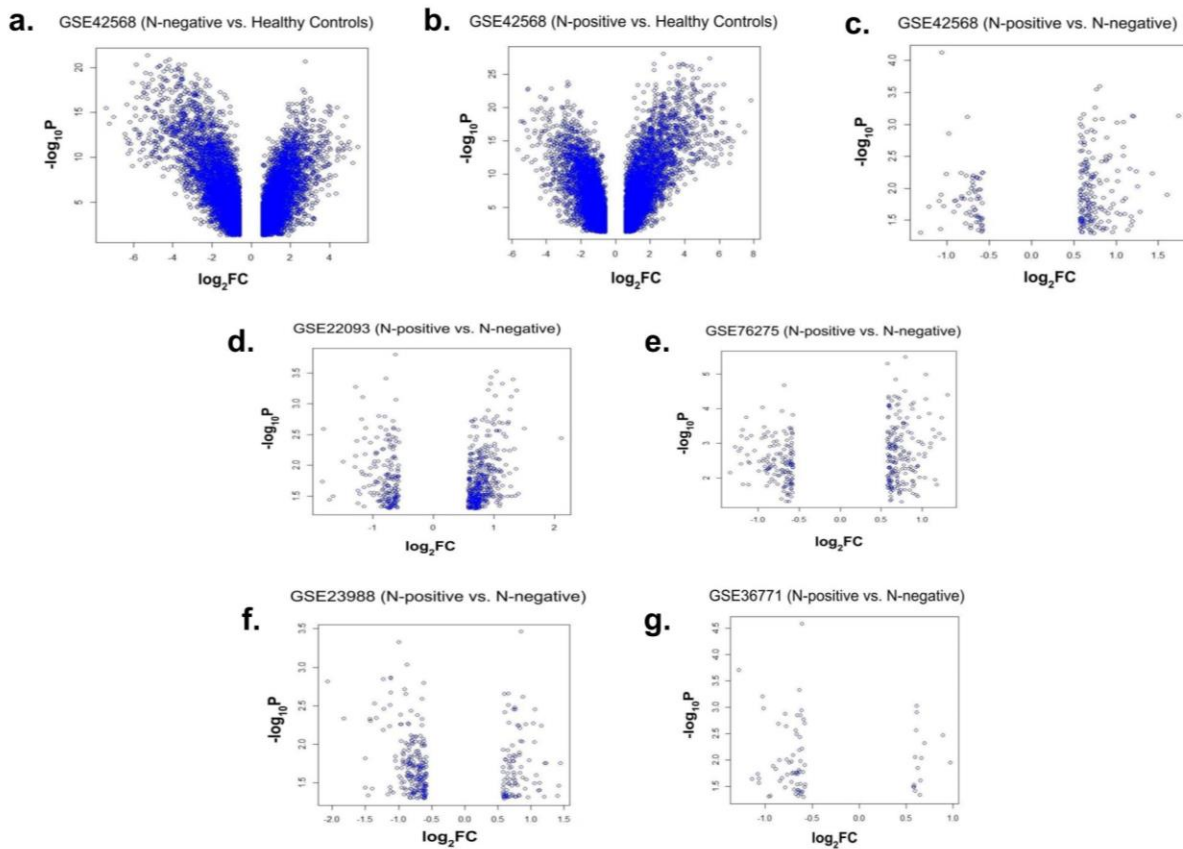
S.No.	Names of genes	Number of methods involved
1.	<b>KRT6B</b>	<b>12</b>
2.	<b>KRT17</b>	<b>11</b>
3.	<b>KRT5</b>	<b>11</b>
4.	KRT23	10
5.	<b>SERPINB5</b>	<b>10</b>
6.	TRIM29	9
7.	<b>KRT81</b>	<b>9</b>
8.	S100A2	9
9.	CDKN2A	8
10.	PRPF31	7
11.	PROM1	6
12.	<b>DSC3</b>	<b>3</b>
13.	CENPI	3
14.	IGHV4-38-2	3
15.	F2R	3
16.	PIP	1
17.	MRPS12	1
18.	PLCB4	1
19.	CWC25	1
20.	P2RY10	1
21.	PQBP1	1

**Table 7:** Common miRNA targeting hub genes

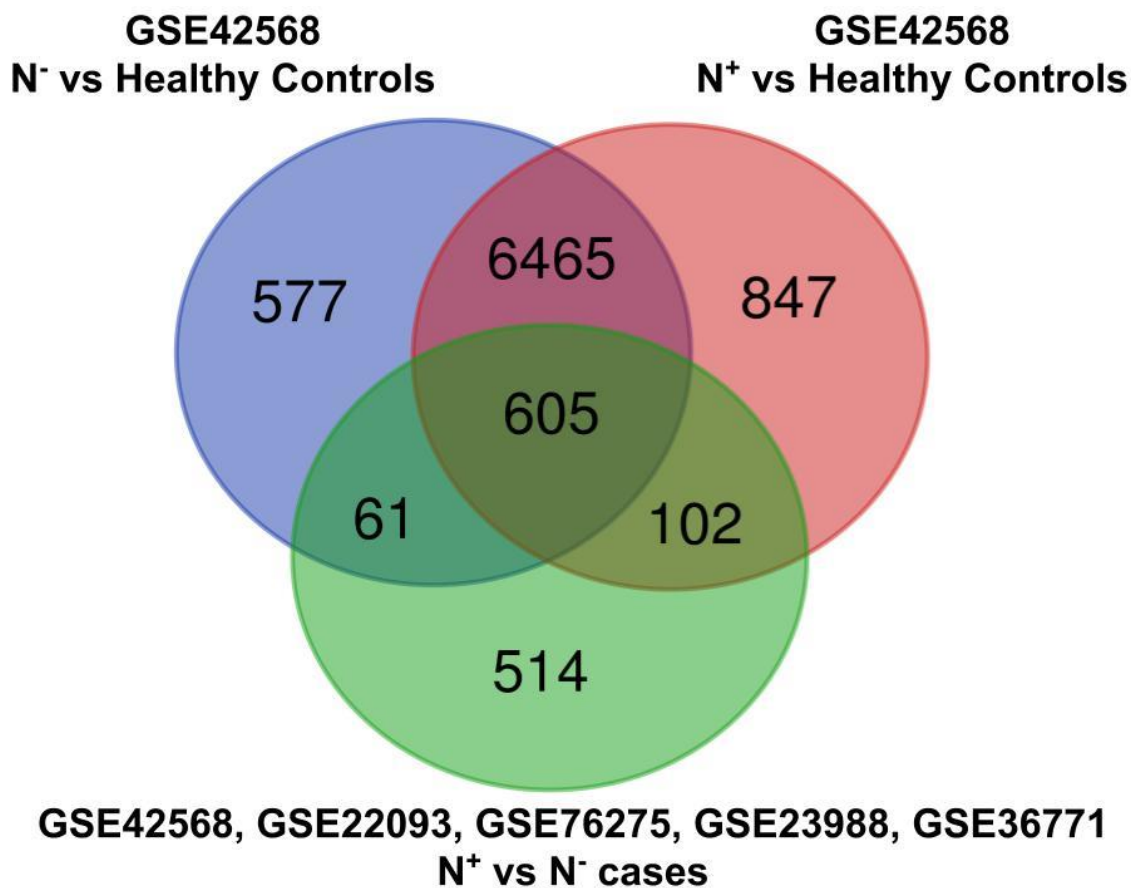
<b>S.No.</b>	<b>Gene</b>	<b>Targeting miRNA</b>
1.	DSC3 KRT17 KRT5 KRT6B SERPINB5	hsa-mir-155-5p
2.	DSC3 KRT17 SERPINB5	hsa-mir-103a-3p
3.	DSC3 KRT6B SERPINB5	hsa-mir-16-5p hsa-let-7b-5p
4.	KRT17 KRT5 SERPINB5	hsa-mir-1343-3p
5.	KRT17 KRT81 SERPINB5	hsa-mir-1-3p hsa-mir-194-5p
6.	DSC3 KRT17	hsa-mir-124-3p
7.	DSC3 KRT81	hsa-mir-129-2-3p hsa-mir-522-5p hsa-mir-27a-3p
8.	DSC3 SERPINB5	hsa-mir-107
9.	KRT17 KRT81	hsa-mir-200b-3p
10.	KRT17 SERPINB5	hsa-mir-429
11.	KRT5 KRT81	hsa-mir-146a-5p
12.	KRT5 SERPINB5	hsa-mir-335-5p hsa-mir-21-5p
13.	KRT6B SERPINB5	hsa-mir-30a-5p



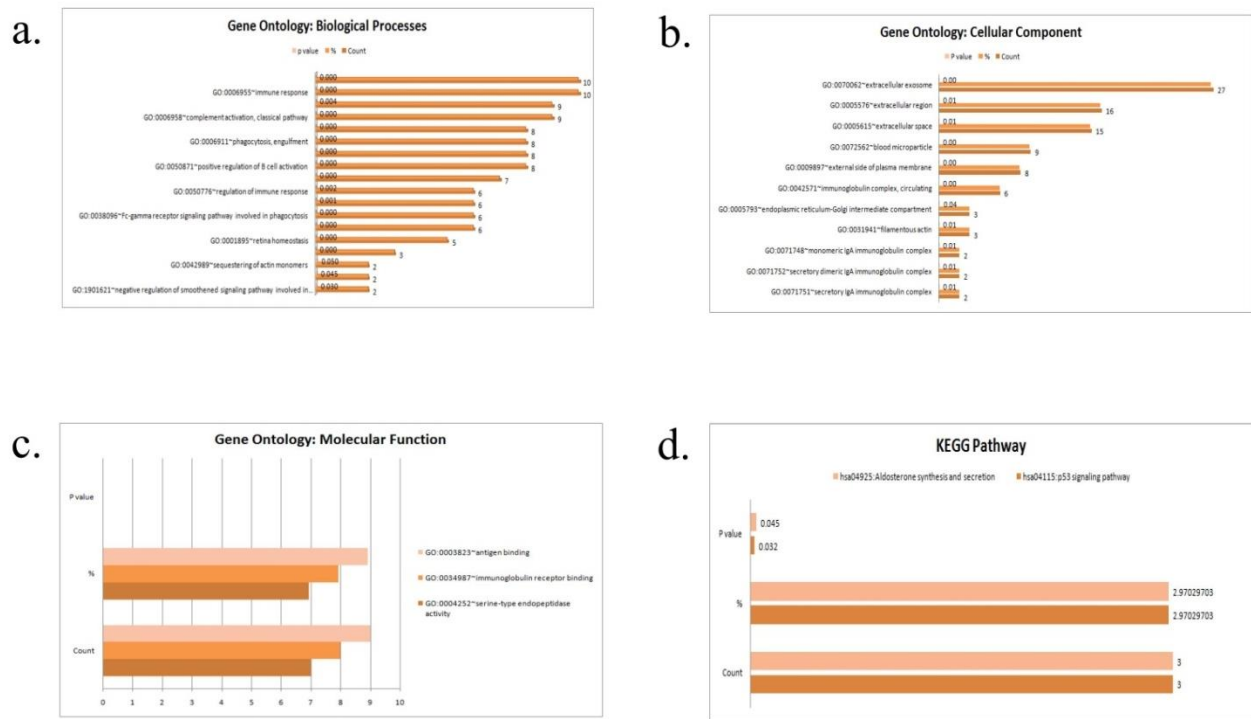
**Figure 1:** A schematic diagram of the workflow



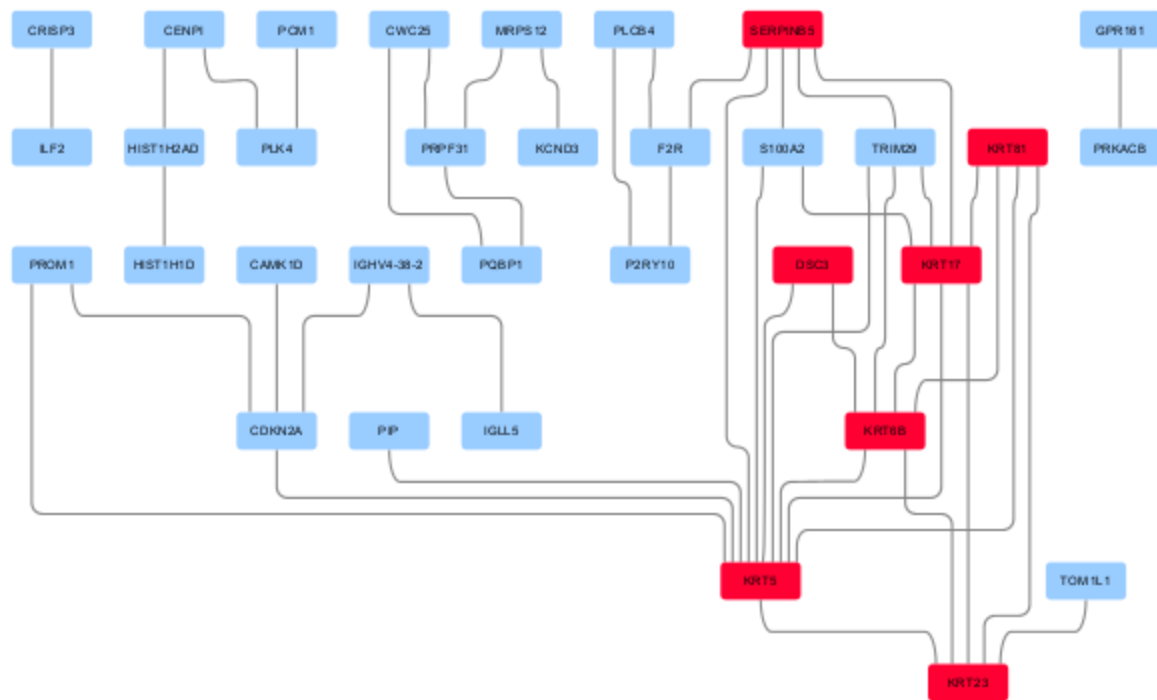
**Figure 2a-g:** Volcano plots for the differentially expressed genes (DEGs) for each dataset



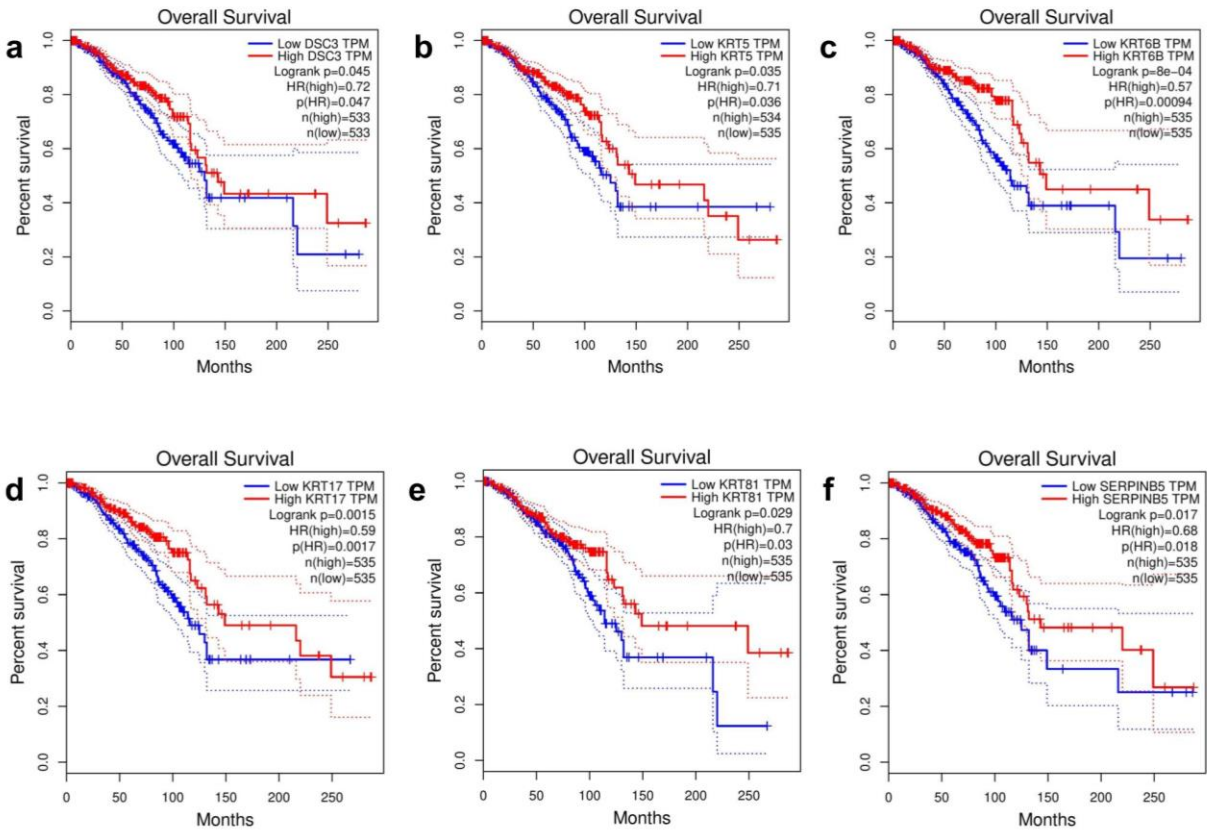
**Figure 3:** Venn diagram depicting the 102 overlapping DEGs for N<sup>+</sup> BC patients



**Figure 4a-d:** Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis by DAVID



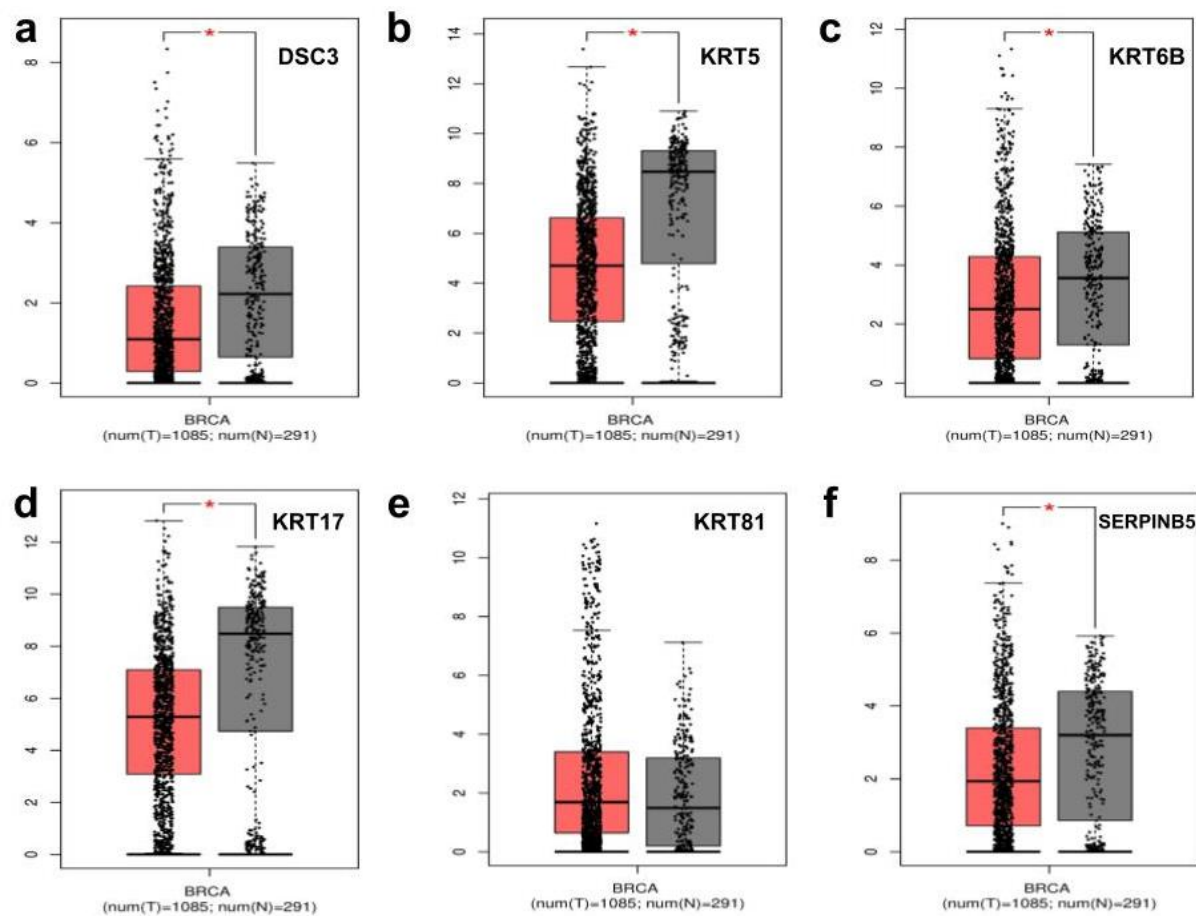
**Figure 5:** The protein-protein interaction (PPI) network for the selected hub genes visualized on Cytoscape



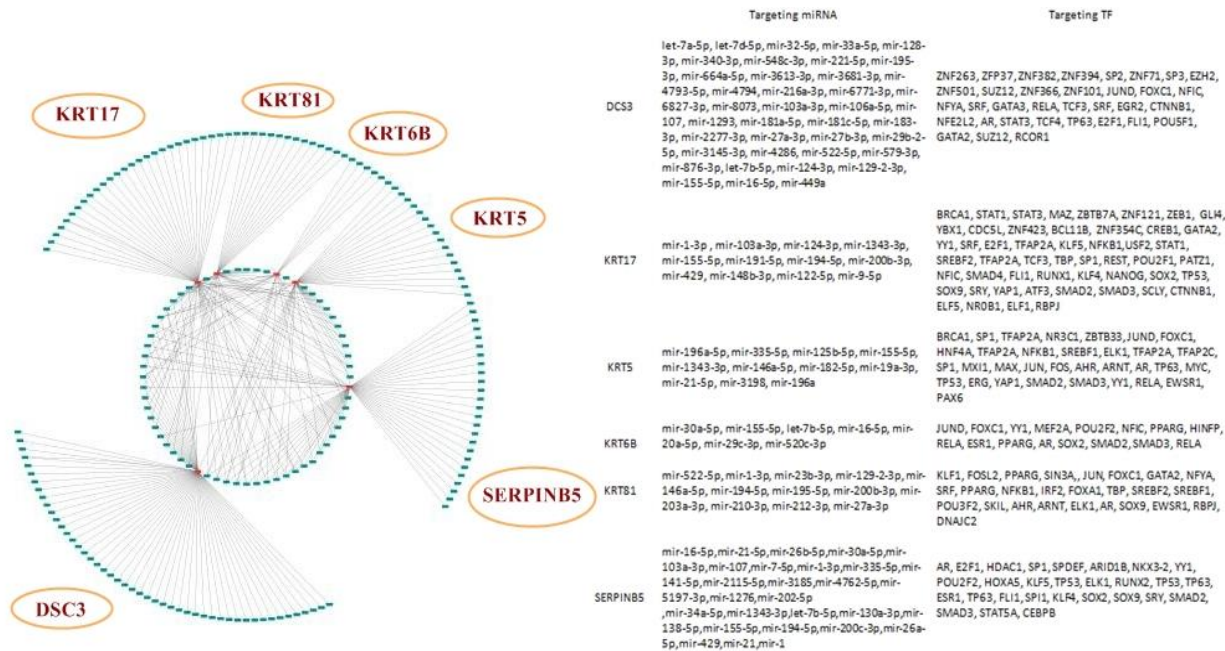
**Figure 6a-f:** Associations of the six candidate hub genes with overall survival in breast cancer.

Log rank  $P<0.05$  is considered as statistically significant





**Figure 7a-f:** Differential expressions of the six candidate hub genes in breast cancer.



**Figure 8:** The hub gene-miRNA-transcription factor regulatory network